Likelihood-based estimation of substructure content from single-wavelength anomalous diffraction (SAD) intensity data

Hatti, Kaushik S.; McCoy, Airlie J.; Read, Randy J.

Published in:
Acta Crystallographica Section D: Structural Biology

DOI:
10.1107/S2059798321004538

Publication date:
2021

STRUCTURAL BIOLOGY
ISSN 2059-7983

Likelihood-based estimation of substructure content from single-wavelength anomalous diffraction (SAD) intensity data

Kaushik S. Hatti,‡ Airlie J. McCoy and Randy J. Read*

Cambridge Institute for Medical Research, Department of Haematology, University of Cambridge, The Keith Peters Building, Hills Road, Cambridge CB2 0XY, United Kingdom. *Correspondence e-mail: rjr27@cam.ac.uk

Received 7 February 2021
Accepted 28 April 2021

Edited by A. Gonzalez, Lund University, Sweden

‡ Current address: Drug Discovery Unit, Wellcome Centre for Anti-Infectives Research, School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, United Kingdom.

Keywords: single-wavelength anomalous diffraction; substructure; likelihood; phasing.

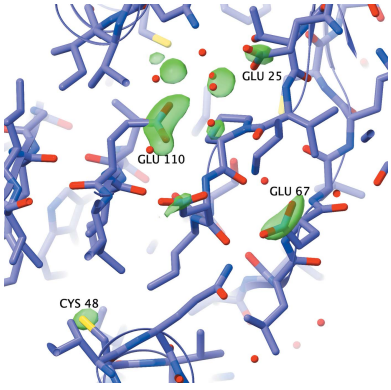
Supporting information: this article has supporting information at journals.iucr.org/d

SAD phasing can be challenging when the signal-to-noise ratio is low. In such cases, having an accurate estimate of the substructure content can determine whether or not the substructure of anomalous scatterer positions can successfully be determined. Here, a likelihood-based target function is proposed to accurately estimate the strength of the anomalous scattering contribution directly from the measured intensities, determining a complex correlation parameter relating the Bijvoet mates as a function of resolution. This gives a novel measure of the intrinsic anomalous signal. The SAD likelihood target function also accounts for correlated errors in the measurement of intensities from Bijvoet mates, which can arise from the effects of radiation damage. When the anomalous signal is assumed to come primarily from a substructure comprising one anomalous scatterer with a known value of f'' and when the protein composition of the crystal is estimated correctly, the refined complex correlation parameters can be interpreted in terms of the atomic content of the primary anomalous scatterer before the substructure is known. The maximum-likelihood estimation of substructure content was tested on a curated database of 357 SAD cases with useful anomalous signal. The prior estimates of substructure content are highly correlated to the content determined by phasing calculations, with a correlation coefficient (on a log–log basis) of 0.72.

1. Introduction

The anomalous differences between Bijvoet pairs of reflections can be exploited for phasing in crystallography. However, the anomalous differences in intensities are generally limited to a few percent in size, and special care needs to be taken in planning the experiment and in collecting and processing the data in order to measure such differences with sufficient precision for successful phasing (Terwilliger *et al.*, 2016*b*). Planning the experiment benefits from estimating the achievable anomalous difference, considering the number of anomalous scatterer sites that might be present and the precision with which the intensities are measured (Terwilliger *et al.*, 2016*a*).

Both *SHELXD* (Schneider & Sheldrick, 2002) and *AutoSol* (Terwilliger *et al.*, 2009), the experimental phasing suite in *Phenix* (Liebschner *et al.*, 2019), require a prior estimate of how many anomalous scatterers are expected in the substructure. The most accurate estimates are obtained when there is a known stoichiometry for an intrinsically bound metal, so that the size of the substructure depends only on the number of copies in the asymmetric unit. For soaking experiments with heavy metals or halides, initial estimates of the number of sites depend on rules of thumb that are typically



based on the number of residues. Even when phasing with intrinsic scatterers such as S atoms in native proteins or with Se atoms in proteins incorporating selenomethionine (SeMet), parts of the chain may be disordered, selenium substitution may be incomplete or radiation damage could reduce their occupancy by the end of the X-ray diffraction data collection.

This work looks at characterizing the data after the experiment has been performed and the data have been processed. Specifically, we are addressing the problem of determining, from the data, the amount of scattering contributed by the anomalous substructure. This provides both an estimate of the size of the actual anomalous differences between Bijvoet pairs and information about the number of sites that is expected in the substructure. The underlying approach is to devise a likelihood target that can be used to determine parameters that quantify the strength of anomalous scattering, considering the effect of errors in measuring intensities of Bijvoet pairs and also the effect of correlations in these errors.

The derivation of the likelihood target starts with understanding how the strength of anomalous scattering affects the sizes of the differences between the true Bijvoet mates, represented through their joint probability distribution. This is developed in Section 2, along with the impact of intensity-measurement errors on the distributions of measured intensities. The manipulation of probability distributions of acentric structure factors is much more straightforward with complex sources of error, so the LLGI approximation is introduced, allowing the effects of scalar errors in intensities to be modelled very well with complex errors.

The likelihood target itself is defined in Section 3, expressed in terms of parameters that depend on the total strength of scattering from the substructure of major anomalous scatterers and other parameters describing the degree to which measurement errors in Bijvoet pairs are correlated. The approximations underlying the likelihood target are validated by showing that they agree very well with exact relationships evaluated by (expensive) numerical integration calculations.

Section 4 develops methods to interpret the adjustable parameters from the likelihood target. It is assumed here that one has knowledge of the strongest anomalous scatterer contained within the crystal, the likely content of the crystal (*i.e.* the number of copies of protein or nucleic acid sequences expected to be found in the asymmetric unit), the wavelength of data collection and the associated scattering factors for anomalous scatterers. Building on this, it is possible to estimate the equivalent number of fully occupied anomalous scatterers, along with their overall B factor, that would be required to explain the differences between Bijvoet mates.

Section 5 describes the collection and curation of a large set of test data and the design of the calculations using these data. Finally, the results of these calculations are outlined in Section 6, evaluating the extent to which the actual substructure content can be predicted from the measured data before a substructure has been determined and refined.

The parameters that are estimated to characterize the SAD intensity data are also required to refine substructure models and obtain phase-probability distributions. This will be investigated in future work, along with ways to assess anomalous signal through measures of information gain and estimates of the log-likelihood-gain score that would be achieved with an ideal substructure model.

2. Intensity-based joint probability distributions for SAD data

To derive probability distributions for measured diffraction data for use in crystallographic likelihood functions, it is necessary to combine the effects of complex differences in the structure factors with those of scalar measurement errors in the intensities. This is further complicated by the fact that the amplitude of the structure factor is related to the square root of the intensity; the true intensity is never negative, but the measured intensity may well be. We have not found a way to derive exact analytical expressions combining these differences. Nonetheless, in our previous work on the LLGI intensity-based likelihood target (Read & McCoy, 2016), we showed that a log-likelihood-gain score that accounts exactly for the effect of Gaussian measurement errors on intensities can be approximated extremely well with a target computed via the Rice function (for the acentric case), in which the intensity and its standard deviation are transformed into an effective amplitude and a Luzzati-style weighting term approximating the effect of the scalar measurement error as an error in the complex plane. Importantly, the effective amplitude and the weighting term are independent of calculated structure factors from a model, so they only need to be determined once. Here, we investigate whether the same approach can be extended to intensity-based iSAD likelihood targets for SAD data, in which there is a pair of correlated intensity measurements for each set of Miller indices. We concentrate on what can be deduced about the joint distribution of the true Bijvoet mates from the corresponding intensity measurements and what this can tell us about the scattering power of the anomalous substructure.

In the following, we make a number of simplifying assumptions.

Firstly, we assume that the intensities (or Bijvoet pairs of intensities) measured for different Miller indices are independent of each other. This is not strictly true, but the correlations arising from the presence of bulk solvent or the existence of noncrystallographic symmetry are much weaker than the strong correlations between Bijvoet pairs.

Secondly, we assume that the phase angles of the individual atomic contributions to structure factors are independent, so that the total structure factors can be considered to arise from a random walk in the complex plane, leading to a complex normal distribution. For this to be true, it is sufficient for the atoms to be randomly distributed in their distances to the Bragg planes associated with any particular reflection; apart from the lowest resolution reflections, it is not necessary to make the much more restrictive assumption that the atoms are randomly distributed throughout the unit cell.

Thirdly, we further assume that this independence extends to the substructure of anomalous scatterers, so that the joint distribution of Bijvoet pairs follows a multivariate complex normal distribution.

Fourthly, we assume that intensity-measurement errors are drawn from a Gaussian distribution, which is independent for reflections with different Miller indices, although there may be a correlation in measurement errors for Bijvoet pairs.

2.1. Joint prior distribution of true Bijvoet mates

To set the stage for characterizing the substructure content, we start by defining the joint distribution of true Bijvoet mates (with no measurement error) in terms of the atomic content of the crystal, divided into the most significant anomalous scatterer (for which a substructure might be determined during the process of phasing) and the rest of the atoms. Note that we are not assuming here that the rest of the atoms lack any anomalous scattering contribution. For instance, in SeMet phasing the S atoms in cysteine residues will make a small but non-negligible contribution to the anomalous differences, even though it is only rarely possible to identify the positions of these atoms during substructure determination.

The Bijvoet mates are described in terms of $\mathbf{F}^{+}$ and the complex conjugate of $\mathbf{F}^{-}$, $\mathbf{F}^{-*}$, because these are highly correlated and have similar phase angles. Individual elements differ in their (in general) complex scattering factor $\mathbf{f}_j$, and each atom will differ in its position $\mathbf{x}_j$, occupancy o_j and displacement parameter B_j, as shown in (1*a*) and (1*b*).

$$\mathbf{F}^{+} = \sum_{j=1}^{N} o_j \exp(-B_j|\mathbf{s}|^2/4)\mathbf{f}_j \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j), \tag{1a}$$

$$\mathbf{F}^{-*} = \sum_{j=1}^{N} o_j \exp(-B_j|\mathbf{s}|^2/4)\mathbf{f}_j^{*} \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j). \tag{1b}$$

In these equations, $\mathbf{h}$ is the vector of Miller indices and $\mathbf{s}$ is the corresponding diffraction vector, the magnitude of which is the inverse of the interplanar spacing. As discussed in our earlier work on SAD phasing (McCoy *et al.*, 2004), the joint distribution of Bijvoet mates takes the form of a multivariate complex normal distribution, which is readily derived by assuming that each atom contributes independently to the total structure factors and considering the atomic parameters to be random variables. [The effects of correlations between atomic contributions arising from translational noncrystallographic symmetry (tNCS) can be addressed by modifying the expected intensity factors in the final equations, as described previously for the case of normal scattering (Read *et al.*, 2013).]. Equation (2*a*) defines the prior joint distribution (before a substructure model is available), where the expected values of the complex Bijvoet mates are zero in the absence of any prior structural knowledge and the Hermitian covariance matrix is defined in (2*b*).

$$\mathrm{p}(\mathbf{F}^{+}, \mathbf{F}^{-*}) = \frac{1}{|\pi\Sigma|}\exp\left[-\begin{pmatrix}\mathbf{F}^{+}\\ \mathbf{F}^{-*}\end{pmatrix}^{H}\Sigma^{-1}\begin{pmatrix}\mathbf{F}^{+}\\ \mathbf{F}^{-*}\end{pmatrix}\right], \tag{2a}$$

where

$$\Sigma = \begin{pmatrix}\langle \mathbf{F}^{+}\mathbf{F}^{+*}\rangle & \langle \mathbf{F}^{+}\mathbf{F}^{-}\rangle \\ \langle \mathbf{F}^{+}\mathbf{F}^{-}\rangle^{*} & \langle \mathbf{F}^{-}\mathbf{F}^{-*}\rangle\end{pmatrix} = \varepsilon\begin{pmatrix}\Sigma_N & \sigma_{FF}\\ \sigma_{FF}^{*} & \Sigma_N\end{pmatrix}. \tag{2b}$$

The diagonal variance term, Σ_N, is simply the scattering power of the crystal defined in terms of the scattering factors in (3*a*), while the off-diagonal covariance element, σ_{FF}, is defined in (3*b*) and ε is the expected intensity factor arising from the statistical effects of crystal symmetry. The superscript H denotes the Hermitian transpose, *i.e.* the transpose of the complex conjugates.

These structure factors are the sums of atomic contributions for N atoms, which will be divided below into the H atoms that could be identified as an anomalous substructure (generally a single primary anomalous scatterer type) and the remaining background (B) atoms that have relatively little anomalous scattering. Depending on the context, intrinsic anomalous scatterers, such as S atoms in cysteine and methionine residues, could either comprise the H atoms or be considered to be part of the B atoms if there is a stronger anomalous scatterer in the crystal. In both cases, the sums can be taken separately over the B and H subsets.

$$\Sigma_N = \sum_{j=1}^{N} o_j^2 \exp(-B_j|\mathbf{s}|^2/2)|\mathbf{f}_j|^2 = \Sigma_B + \Sigma_H, \tag{3a}$$

$$\sigma_{FF} = \sum_{j=1}^{N} o_j^2 \exp(-B_j|\mathbf{s}|^2/2)\mathbf{f}_j^2 = \sigma_{BB} + \sigma_{HH}. \tag{3b}$$

The scattering factor can be expressed as $\mathbf{f}_j = (f_0 + f') + if''$, which is a function of both wavelength and resolution. Note that the wavelength-dependent correction terms f' and f'' are essentially independent of resolution as they arise from inner-shell electrons that can be considered to be point scatterers at the relevant resolutions. The wavelength-independent form factor f_0 provides the resolution dependence. For (3*a*) and (3*b*), we can expand the scattering factor terms to obtain

$$|\mathbf{f}_j|^2 = (f_0 + f')^2 + f''^2$$

and

$$\mathbf{f}_j^2 = (f_0 + f')^2 - f''^2 + 2i(f_0 + f')f''.$$

The structure factors can be normalized to give E-values with a mean-square value of 1 by dividing them by the square root of their expected intensities, $\varepsilon\Sigma_N$. In the joint distribution of E-values, there is just a single complex correlation parameter, ρ_{FF}:

$$\mathrm{p}(\mathbf{E}^{+}, \mathbf{E}^{-*}) = \frac{1}{|\pi\Sigma|}\exp\left[-\begin{pmatrix}\mathbf{E}^{+}\\ \mathbf{E}^{-*}\end{pmatrix}^{H}\Sigma^{-1}\begin{pmatrix}\mathbf{E}^{+}\\ \mathbf{E}^{-*}\end{pmatrix}\right], \tag{4a}$$

$$\text{where } \Sigma = \begin{pmatrix}1 & \rho_{FF}\\ \rho_{FF}^{*} & 1\end{pmatrix} \tag{4b}$$

$$\text{and } \rho_{FF} = \frac{\sigma_{FF}}{\Sigma_N}. \tag{4c}$$

Because the Bijvoet pairs are highly correlated, values of ρ_{FF} in practice have magnitudes of only slightly less than one. The deviation from one tends to increase with resolution, because f'' is effectively independent of resolution, whereas the real parts of the scattering factors decrease with resolution.

2.2. Correlated measurement errors in measured Bijvoet mates

In the LLGI approach to accounting for the effect of measurement error, the intensity and its standard deviation are transformed into an effective amplitude F_e (or E_e for normalized data) and a Luzzati-style weighting factor D_O that, together in a Rice probability function, give an excellent approximation to the posterior probability of the true amplitude given the intensity. In the related iSAD approach to an intensity-based likelihood function proposed here, both members of the Bijvoet pair are transformed in the same way.

As demonstrated below, this approach is well justified when the measurement errors in the Bijvoet mates are uncorrelated, but requires some elaboration when they are correlated. As discussed by Garcia-Bonete & Katona (2019), time-dependent effects on the measured intensities, such as radiation damage, can lead to correlations between the errors of mean intensity measurements, and there is evidence of such correlations in some of the data sets that we have examined (discussed below). Correlations in measurement errors can be accounted for by assuming that the errors are drawn from a bivariate normal distribution in which the individual variances are obtained from the data-processing analysis but in which a nonzero correlation is present. For simplicity of notation, we use Z to represent the square of an E-value (or, equivalently, a normalized intensity). A joint probability distribution for the effect of correlated measurement errors on the observed normalized intensities is given in (5).

$$\begin{aligned} p(Z_O^+, Z_O^-; Z^+, Z^-) = {} & \frac{1}{2\pi[(1-\rho_\pm^2)\sigma_{Z_O^+}^2\sigma_{Z_O^-}^2]^{1/2}} \\ & \times \exp\left[-\frac{(Z_O^+ - Z^+)^2}{2\sigma_{Z_O^+}^2(1-\rho_\pm^2)} - \frac{(Z_O^- - Z^-)^2}{2\sigma_{Z_O^-}^2(1-\rho_\pm^2)} \right. \\ & \left. + \frac{\rho_\pm(Z_O^+ - Z^+)(Z_O^- - Z^-)}{\sigma_{Z_O^+}\sigma_{Z_O^-}(1-\rho_\pm^2)} \right], \end{aligned} \tag{5}$$

where Z^+ and Z^- are the true values of the normalized intensities, Z_O^+ and Z_O^- are their respective observed values, $\sigma_{Z_O^+}$ and $\sigma_{Z_O^-}$ are the respective standard deviations of the measurements and $\rho_\pm$ is the correlation coefficient between the measurement errors.

It seems reasonable to conjecture that the effect of this correlation on the iSAD approximation can be modelled by assuming that the implied complex errors in the structure factors are correlated to the same degree as the real errors in the corresponding measured intensities. In the iSAD approximation (as in the LLGI approximation), the effective normalized amplitude arises from a complex structure factor that is obtained by adding a complex normal error to the down-weighted true structure factor, as given in (6).

$$\mathbf{E}_e^+ = D_O^+\mathbf{E}^+ + \mathbf{\Delta}^+, \tag{6a}$$

$$\text{where } \langle|\mathbf{\Delta}^+|^2\rangle = 1 - D_O^{+2}. \tag{6b}$$

In this approximation, D_O^+ plays the role of a complex correlation between the true $\mathbf{E}^+$ and the phased effective amplitude $\mathbf{E}_e^+$. Note that, because of the D_O^+ weight on $\mathbf{E}^+$, the expected value of $(E_e^+)^2$ is one. Equivalent expressions apply to the Bijvoet mate. The assumption that the complex errors are correlated to each other, with a complex correlation coefficient that has a magnitude equal to $\rho_\pm$, allows us to determine the complex correlation between $\mathbf{E}_e^+$ and $\mathbf{E}_e^{-*}$, defined as $\rho_{FF,obs}$, by analogy to the complex correlation ρ_{FF} between the corresponding true values $\mathbf{E}^+$ and $\mathbf{E}^-$. This is shown in (7), where we assume that the complex errors are uncorrelated with the true weighted structure factors, so that cross-terms such as $D_O^+\mathbf{E}^+\mathbf{\Delta}^-$ disappear.

$$\begin{aligned} \rho_{FF,obs} &= \langle \mathbf{E}_e^+\mathbf{E}_e^- \rangle = \langle (D_O^+\mathbf{E}^+ + \mathbf{\Delta}^+)(D_O^-\mathbf{E}^- + \mathbf{\Delta}^-) \rangle \\ &= D_O^+D_O^-\rho_{FF} + \langle \mathbf{\Delta}^+\mathbf{\Delta}^- \rangle, \\ \rho_{FF,obs} &= D_O^+D_O^-\rho_{FF} + \rho_\pm[(1-D_O^{+2})(1-D_O^{-2})]^{1/2}. \end{aligned} \tag{7}$$

For simplicity (also justified by the consideration that the implied complex error is effectively modelling the error in the amplitude, *i.e.* the error parallel to the structure factor), we will assume that $\rho_\pm$ (and thus $\rho_{FF,obs}$) has the same phase as ρ_{FF}. In any event, in the situations considered here only the absolute value of $\rho_{FF,obs}$ influences the outcome, although the phase of the complex correlation will influence phasing calculations when a substructure model is considered in future work.

By analogy to (4), the joint distribution of the phased effective amplitudes is defined in (8).

$$p(\mathbf{E}_e^+, \mathbf{E}_e^{-*}) = \frac{1}{|\pi\Sigma|}\exp\left[-\begin{pmatrix} \mathbf{E}_e^+ \\ \mathbf{E}_e^{-*} \end{pmatrix}^H \Sigma^{-1} \begin{pmatrix} \mathbf{E}_e^+ \\ \mathbf{E}_e^{-*} \end{pmatrix} \right], \tag{8a}$$

$$\text{where } \Sigma = \begin{pmatrix} 1 & \rho_{FF,obs} \\ \rho_{FF,obs} & 1 \end{pmatrix}. \tag{8b}$$

3. The data likelihood target: joint distribution of effective amplitudes

The probability distribution in (8) relates structure factors, but the measured data are intensities with unknown phases, which have been transformed into the effective amplitudes in this equation. The phases in (8) can be integrated out to obtain a likelihood function that depends only on the effective amplitudes, given in (9).

$$\begin{aligned} p(E_e^+, E_e^-) = {} & \frac{4E_e^+E_e^-}{1-|\rho_{FF,obs}|^2}\exp\left(-\frac{E_e^{+2} + E_e^{-2}}{1-|\rho_{FF,obs}|^2} \right) \\ & \times I_0\left(\frac{2|\rho_{FF,obs}|E_e^+E_e^-}{1-|\rho_{FF,obs}|^2} \right). \end{aligned} \tag{9}$$

Note that there is only a single parameter to describe the variance of this distribution, $\rho_{FF,\mathrm{obs}}$. However, $\rho_{FF,\mathrm{obs}}$ is itself a function of D_{O}^{+} and D_{O}^{-}, which are fixed values obtained in the calculation of the effective amplitudes, and of the adjustable parameters ρ_{FF} and $\rho_{\pm}$. As discussed above, $\rho_{FF,\mathrm{obs}}$ can be treated as a scalar (as well as the underlying ρ_{FF}) in this context, because any phase component has no effect on the likelihood in the absence of a substructure model.

This likelihood function, which is the main focus of this work, can be used for two purposes. Firstly, the adjustable variance parameters can be refined to characterize the data in terms of the strength of the anomalous scattering ($|\rho_{FF}|$) and potentially the degree to which the measurement errors are correlated ($\rho_{\pm}$), if this parameter is not available from an analysis during the merging step of data processing. Secondly, it provides the likelihood score for a null hypothesis in phasing, *i.e.* the baseline for a log-likelihood gain (LLG) when a substructure model is available. In other words, it can play an equivalent role to the Wilson distribution (Wilson, 1949) in the LLG used for purely real scattering in molecular replacement or refinement. Here, we will explore the uses of this likelihood function to characterize the data, particularly to estimate the substructure content.

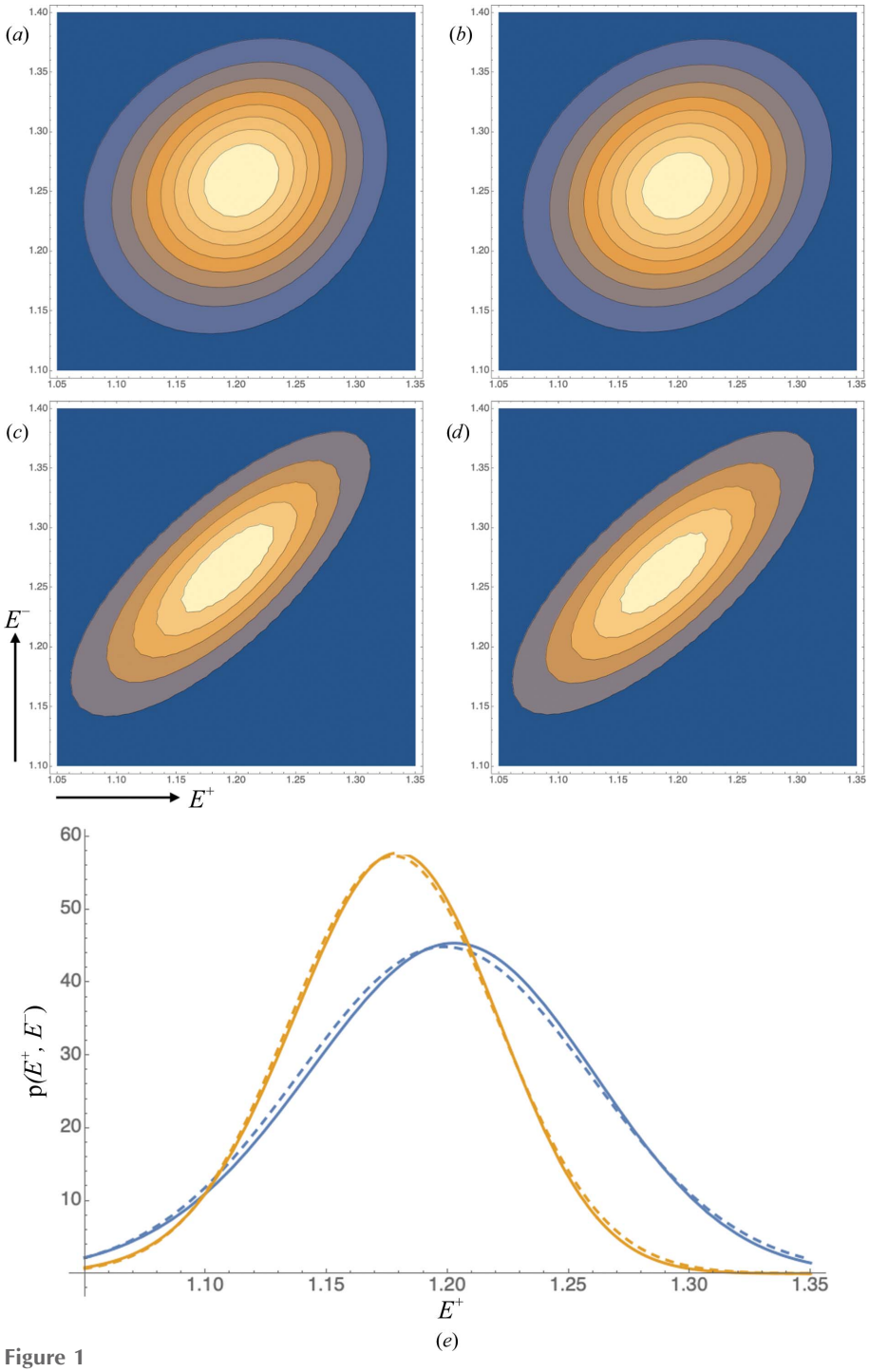


Figure 1
Comparison of exact and approximate probability distributions for the true normalized amplitudes conditional on the observed intensities. (*a*) Contour plot illustrating $p(E^{+}, E^{-}; Z_{\mathrm{O}}^{+}, Z_{\mathrm{O}}^{-})$ for $\rho_{\pm} = 0$. (*b*) Contour plot illustrating $p(E^{+}, E^{-}; E_{\mathrm{e}}^{+}, E_{\mathrm{e}}^{-})$ for $\rho_{\pm} = 0$. (*c*) Contour plot illustrating $p(E^{+}, E^{-}; Z_{\mathrm{O}}^{+}, Z_{\mathrm{O}}^{-})$ for $\rho_{\pm} = 0.75$. (*d*) Contour plot illustrating $p(E^{+}, E^{-}; E_{\mathrm{e}}^{+}, E_{\mathrm{e}}^{-})$ for $\rho_{\pm} = 0.75$. (*e*) Slices through the joint probability distributions at $E^{-} = 1.25$ for the cases shown in (*a*) (solid blue line), (*b*) (dashed blue line), (*c*) (solid orange line) and (*d*) (dashed orange line).

3.1. Validation of the iSAD approximation

To verify that it is appropriate, firstly to construct the iSAD approximation by transforming the observed intensities independently into effective amplitudes and D_{O} factors and secondly to assume that the same correlation parameter $\rho_{\pm}$ can be used to model the effect of correlated measurement error, we have followed the approach used in validating the LLGI target (Read & McCoy, 2016) by comparing the conditional probabilities of the true amplitudes given the observations, obtained either with the exact treatment or with the iSAD approximation.

The gold standard for the comparison is the joint conditional probability distribution for the true amplitudes given the observed normalized intensities, denoted as Z-values $[p(E^{+}, E^{-}; Z_{\mathrm{O}}^{+}, Z_{\mathrm{O}}^{-})]$, derived by following the propagation of errors and using numerical integration, giving (19) in Appendix *A*. The corresponding joint conditional distribution, given the effective amplitudes from the iSAD approximation $[p(E^{+}, E^{-}; E_{\mathrm{e}}^{+}, E_{\mathrm{e}}^{-})]$, is provided as (23) in Appendix *B*.

Fig. 1 provides two comparisons of these joint probability distributions in a

calculation modelled on SeMet phasing where both the intrinsic anomalous signal and the measurement error are significant. In one case the measurement errors are assumed to be independent, whereas in the second case the errors are assumed to be highly correlated, with $\rho_{\pm} = 0.75$. The exact distribution and the iSAD approximation are indeed very similar in both cases, while the introduction of correlated errors has a profound effect on the distributions. Similar results were obtained in other calculations where the level of anomalous signal, the measurement error and the correlation of measurement error have been varied (not shown), justifying the use of this approach.

4. Maximum-likelihood estimation of substructure content

When phasing with intrinsic anomalous scatterers, such as Se atoms in SeMet constructs or S atoms in native proteins, one has reasonable prior knowledge of the atomic composition of the crystal. Even in this favourable case, there is uncertainty about the degree to which the expected sites are ordered and potential uncertainty about the occupancy of Se sites because of variable SeMet incorporation. When soaking with heavy-atom compounds, halides or other derivatives, only a rough guess can be made in advance about the degree of substitution. Refinement of the variance parameters in a log-likelihood function based on (9) should enable a reduction of the uncertainty in the substructure content relative to other atoms in the crystal. This will be useful in characterizing the phasing signal as well as in judging the difficulty of substructure determination.

There is a direct relationship between $|\rho_{FF}|$ and relative substructure content if we treat the scattering power of only one primary type of anomalous scatterer as unknown. The anomalous scatterer content can be placed on an absolute scale if the number of copies of the protein in the crystal can be deduced from the Matthews volume (Matthews, 1968). Equation (10) is a simple consequence of (3) and (4), given that the primary anomalous scatterer (H) atoms share the same scattering factor, denoted $\mathbf{f}_H$ here.

$$|\rho_{FF}| = \frac{|\sigma_{BB} + \sigma_{HH}|}{\Sigma_B + \Sigma_H} = \frac{|\sigma_{BB} + x_H \mathbf{f}_H^2 \exp(-B_{\rm W}|\mathbf{s}|^2/2)|}{\Sigma_B + x_H |\mathbf{f}_H|^2 \exp(-B_{\rm W}|\mathbf{s}|^2/2)}, \quad (10a)$$

$$\text{where } x_H = \sum_{j=B+1}^{N} o_j^2 \exp(-\Delta B_j |\mathbf{s}|^2/2). \quad (10b)$$

In (10a) the overall Wilson B factor ($B_{\rm W}$) has been factored out of the primary anomalous scatterer contributions, leaving the individual atomic differences (ΔB_j) in (10b). For the substructure content analysis, these equations are simplified by factoring out the overall Wilson B factor from all sums, approximating the B (other background) atoms as sharing the same overall B factor, and approximating the H atoms as sharing the same ΔB_H relative to the B factor of the B atoms. These approximations give (11).

$$|\rho_{FF}| \simeq \frac{|\sigma_{BB,0} + \sigma_{HH,0}|}{\Sigma_{B,0} + \Sigma_{H,0}} = \frac{|\sigma_{BB,0} + x_H \mathbf{f}_H^2|}{\Sigma_{B,0} + x_H |\mathbf{f}_H|^2}, \quad (11a)$$

$$\text{where } \sigma_{BB,0} = \sum_{j=1}^{B} o_j^2 \mathbf{f}_j^2, \quad (11b)$$

$$\Sigma_{B,0} = \sum_{j=1}^{B} o_j^2 |\mathbf{f}_j|^2, \quad (11c)$$

$$x_H = \exp(-\Delta B_H |\mathbf{s}|^2/2) \sum_{j=B+1}^{N} o_j^2 = n_H \exp(-\Delta B_H |\mathbf{s}|^2/2). \quad (11d)$$

In (11d), n_H is the equivalent number of fully occupied atoms with the same total scattering power as the substructure, (which is weighted by the sum of occupancies squared); this is not necessarily and indeed is not usually an integer.

To convert $|\rho_{FF}|$ for a resolution shell to a value of x_H, (11a) is solved for x_H by transforming it into a quadratic expression in x_H, shown in (12).

$$|\rho_{FF}|^2 (\Sigma_{B,0} + x_H |\mathbf{f}_H|^2)^2 - |\sigma_{BB,0} + x_H \mathbf{f}_H^2|^2 = 0, \quad (12a)$$

$$a x_H^2 + b x_H + c = 0, \quad (12b)$$

$$\text{where } \mathbf{f}_H = f_H + i f_H'', \quad (12c)$$

$$a = (1 - |\rho_{FF}|^2) |\mathbf{f}_H|^4, \quad (12d)$$

$$b = 4 f_H f_H'' \text{Im}(\sigma_{BB,0}) + 2 f_H^2 [\text{Re}(\sigma_{BB,0}) - |\rho_{FF}|^2 \Sigma_{B,0}] - 2 f_H''^2 [\text{Re}(\sigma_{BB,0}) + |\rho_{FF}|^2 \sigma_{B,0}], \quad (12e)$$

$$c = \text{Re}(\sigma_{BB,0})^2 + \text{Im}(\sigma_{BB,0})^2 - (|\rho_{FF}| \Sigma_{B,0})^2. \quad (12f)$$

There are in general two solutions to the quadratic, as illustrated in Fig. 2. In the current implementation, the solution corresponding to a smaller substructure is chosen, although if a prior probability distribution for the substructure size were provided the two solutions could be assigned relative posterior probabilities.

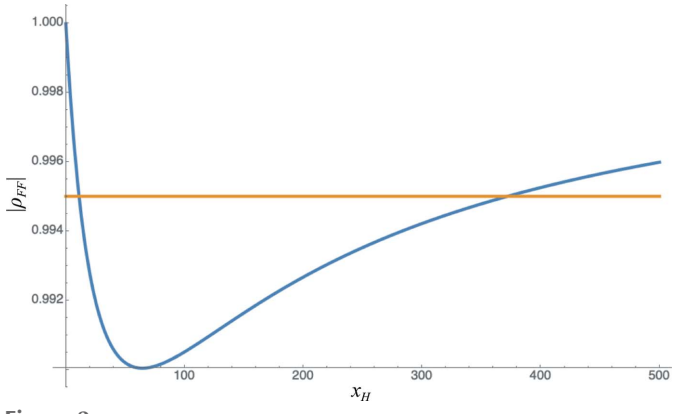


Figure 2
Complex correlation as a function of substructure composition. The calculated magnitude of the complex correlation, $|\rho_{FF}|$, is shown in blue as a function of the assumed number of Se atoms in the asymmetric unit (computed against a background of 1000 C, N or O atoms and ten S atoms). Intersections with the horizontal orange line illustrate that a refined value of 0.995 for $|\rho_{FF}|$ is consistent with either about 11 Se atoms or 370 Se atoms. The minimum value of $|\rho_{FF}|$ consistent with the assumed background composition and nature of the primary anomalous scatterer is about 0.990.

One approach that has been tested is to use the resulting x_H values for resolution bins to estimate values of n_H and ΔB_H by transforming (11d) into (13) and fitting a least-squares line.

$$\ln(x_H) = \ln(n_H) - \Delta B_H |\mathbf{s}|^2/2. \tag{13}$$

However, we have found that a slightly better stability is obtained with an alternative approach. The target function given in (14) is minimized, starting from a grid search varying n_H and ΔB_H over a range of values consistent with the x_H estimates obtained from the refined $|\rho_{FF}|$ values.

$$T = \left(\frac{\Delta B_H}{\sigma_{B_H}}\right)^2 + \sum_{j=1}^{N_{\rm bins}} \frac{\delta_j^2}{2(1+\kappa\delta_j^2)}, \tag{14a}$$

$$\text{where } \delta_j^2 = \frac{1}{\sigma^2_{|\rho_{FF}|_j}}(|\rho_{FF}|_j - |\rho_{FF,\rm calc}|_j)^2. \tag{14b}$$

The first term in T restrains ΔB_H to zero, with a standard deviation typically set to 5 Å^2. The factor κ, which is typically set to 0.1, controls the damping of the robust Geman–McClure loss function comprising the second term. The calculated values of $|\rho_{FF}|$ are computed using (11). The standard deviations for $|\rho_{FF}|$ values are obtained from the inverse of the second-derivative (Hessian) matrix of the likelihood target computed for the optimized parameters.

4.1. Strategy for the refinement of variance parameters

Refinement of the $|\rho_{FF}|$ and $\rho_\pm$ parameters to maximize the likelihood function based on (9) is implemented in the SCA (substructure content analysis) mode of *Phasertng*, which is under development (McCoy *et al.*, 2021). In the current implementation these parameters are refined in resolution bins, with a minimum of 500 reflections per bin. Two refinement macrocycles are carried out. In both macrocycles the bin values for $\rho_\pm$ are constrained to lie in the range 0–0.9, with a weak quadratic restraint towards the value of 0 (standard deviation of 0.5) so that error correlations are inferred only when required to explain the data. In addition, a quadratic smoothness restraint penalizes $\rho_\pm$ values that differ from the value computed from the line connecting the two nearest neighbours (standard deviation of 0.025). This is similar to an approach used to stabilize the refinement of $\sigma_{\rm A}$ values for maximum-likelihood refinement when they are evaluated using just the cross-validation data (Pannu & Read, 1996). In the first macrocycle, the bin values for $|\rho_{FF}|$ are constrained to lie in the range 0–1 while being otherwise unrestrained. At the end of this macrocycle, values of n_H and ΔB_H are estimated from the bin values for $|\rho_{FF}|$ as discussed above. Some values of $|\rho_{FF}|$ are too low to be achieved with any value of x_H for a given anomalous scatterer, as shown in Fig. 2. Resolution bins violating this constraint are ignored in the determination of n_H and ΔB_H, and their values for $|\rho_{FF}|$ are reset to those computed from the values of n_H and ΔB_H estimated from all of the data. This situation generally arises near the resolution limit, when the anomalous signal is very small relative to the noise.

For the second macrocycle, the estimated values of n_H and ΔB_H are used to determine target values for x_H, and thus $|\rho_{FF}|$, for each resolution shell. Loose restraints for $|\rho_{FF}|$ are applied to smooth the curve as a function of resolution, with the standard deviation being determined by the change in $|\rho_{FF}|$ that would change x_H by a factor of 1.5. This can stabilize refinement in cases with weak signal to noise, but has relatively little effect in most cases. In addition, $|\rho_{FF}|$ in each bin is constrained in this macrocycle to lie between the minimum that can be achieved with any value for x_H and the maximum possible value, corresponding to $x_H = 0$.

5. Methods

5.1. Collecting and curating test data

The method to determine substructure content was tested on a database of SAD data sets provided by collaborators or downloaded from the Worldwide Protein Data Bank (wwPDB; Berman *et al.*, 2000). 124 data sets were kindly provided by Zbigniew Dauter, most of which have been discussed earlier (Banumathi *et al.*, 2004; Dauter *et al.*, 2002; Wang *et al.*, 2006). 162 data sets (which include MAD data sets split into individual wavelengths and considered as SAD data sets) were collated by Tom Terwilliger from JCSG experiments and have been discussed earlier (Bunkóczi *et al.*, 2015).

The majority of the data sets in the database were downloaded directly from the wwPDB. The advanced search option of the RCSB PDB was used to perform queries. A list of PDB entries was collected which had a 'Structure Determination Method' record containing the word 'SAD' and a 'Citation' record, and for which experimental data including Bijvoet pairs had been deposited. Data extending to poorer than 4 Å resolution and structures possessing tNCS were excluded. This list was split into three categories.

(i) Soaking experiments, comprising structures determined with any halides, heavy metals, noble gases or other elements from derivatives commonly used in phasing experiments.

(ii) SeMet experiments, comprising structures containing Se atoms (in order for these not to dominate the database SeMet structures were restricted to entries deposited after 1 January 2018).

(iii) Sulfur SAD phasing experiments, which were identified by examining PDB entries that provide Bijvoet pairs but do not contain any atoms heavier than S.

For each entry, the *Phenix* package *phenix.fetch_pdb* command with the argument `--mtz` was used to download the model, sequence and structure factors, and to convert structure factors from cif to MTZ file format. The values for wavelength, cell dimensions, resolution and space group were verified with the associated publications, and any inconsistent data were removed from the list. Each data set was associated with the element type expected to contribute most strongly to the anomalous signal, denoted the primary anomalous scatterer. A total of 536 data sets were selected initially. We were surprised to note that none of these are affected by twinning, an observation that highlights the difficulty that current phasing methods have with such data.

The data sets were screened for the presence of at least minimal anomalous signal during the initial step to generate reference substructures using the MR-SAD protocol (discussed below). Several data sets had so little anomalous signal that no anomalous scatterers could be detected. A significant number of other data sets had such poor anomalous signal that only a small fraction of atoms in the substructure were placed correctly. These data sets were omitted from the subsequent analysis, leaving 382 of the original 536. It seems likely that many of these data sets are in fact native data for structures that were solved by SAD phasing using separate data that were not deposited. For a small additional number of data sets that were omitted, the reported wavelength was incompatible with the strength of the anomalous signal. This left 357 data sets in the curated database.

5.2. Generating reference substructures

Reference substructures were generated using the MR-SAD protocol available in *Phaser* (Read & McCoy, 2011). To be consistent in the use of structure-factor amplitudes (needed for the current version of *Phaser*, which does not work with intensity data), deposited intensity data (whenever available) were converted to structure-factor amplitudes ($|\mathbf{F}|$) and their estimated standard deviations for the MR-SAD step using the *phenix.french_wilson* tool (Liebschner *et al.*, 2019). For data sets with only deposited structure-factor amplitudes, these were converted to approximate intensity measurements as described for the LLGI target (Read & McCoy, 2016) for the substructure content analysis step. In the MR-SAD protocol, the deposited atomic model of the protein is used as a starting model for phasing, but is treated as being composed of purely real scatterers. In the approach used here, anomalously scattering centres were found using SAD log-likelihood-gradient maps (McCoy & Read, 2010) to search for purely imaginary scatterers, since the real scattering at each centre was already accounted for in the deposited model used for phasing.

Purely imaginary scatterers found in the MR-SAD step were replaced with the atom type of the corresponding atom in the deposited structure to give the anomalous substructure, annotated using *phenix.emma* (Grosse-Kunstleve & Adams, 2003) to identify atoms that superimpose within a distance threshold. The parameters of the anomalous substructure were then refined against the data, without altering the substructure with log-likelihood-gradient completion (Read & McCoy, 2011). The refined f'' for the primary anomalous scatterer and, for each anomalous scatterer type identified, the number of sites and the sum of the squared occupancies of sites, were stored in the database. The total scattering power of the anomalous substructure was evaluated in terms of the equivalent number of fully occupied primary anomalous scatterers, which was calculated as the sum of squared occupancies for each atom type weighted by the square of the ratio of the f'' for that anomalous scatterer type and the f'' of the primary anomalous scatterer. This approximation assumes that the contribution of any secondary anomalous scatterers, if present, is dominated by their imaginary contribution, and that differences among atom types in the ratio of real to imaginary scattering are less important. The quality of SAD phasing was assessed by computing the correlation between the experimentally phased map and density generated from the deposited model using *phenix.get_cc_mtz_pdb*.

5.3. Choice of refined f'' over theoretical f'' for estimating anomalous signal

Many of the test data sets have been measured at a wavelength near the absorption edge of the primary anomalous scatterer, where the f'' changes rapidly. For these data sets, the f'' for the primary scatterer was refined as part of substructure refinement and phasing. Values of f'' obtained from table lookup can have significant errors: the tabulated values do not account for the effects of the chemical environment (Evans & Pettifer, 2001) and the wavelength may not be known precisely because of errors in monochromator calibration (Ruslan Sanishvili, personal communication). It is best to obtain prior estimates of f'' from a fluorescence scan of the crystal at the beamline (Evans & Pettifer, 2001), but in this study we do not have access to fluorescence-scan data for the test data sets. For the data sets collected near the absorption edge, we have therefore used the refined f'' values for the primary anomalous scatterer obtained during refinement and phasing with the reference substructure. We expect the refined f'' to be a better estimate of the true f'' than the value from a table lookup, but there will be random errors. In the refinement, the f'' value for an anomalous scatterer type and the overall occupancies of the individual atoms will be correlated, with both changing the imaginary terms in calculated structure factors but differing in how they affect the relative contributions of the real and imaginary terms as a function of resolution; how well these effects are decoupled will depend on the precision of the data. There may also be systematic errors. For instance, if there is a mixed substructure and some atom types are incorrectly identified, the refined f'' values will reflect a compromise between the relative real and imaginary scattering of the different atom types.

5.4. Preparation of data for substructure content analysis (SCA)

The diffraction data were processed using the *Phasertng.xtricorder* module of *Phasertng* (McCoy *et al.*, 2021). *Phasertng.xtricorder* carries out a series of data analyses to detect and correct for the statistical effects of anisotropy, tNCS and twinning, although none of the data included in this study were affected by either tNCS or twinning. The data were scaled and used for maximum-likelihood estimation of substructure content, which is carried out within *Phasertng.xtricorder* when the data include Bijvoet pairs. The known protein composition of the crystal was used when scaling the data; if an incorrect composition were used, the intensity scaling and therefore the estimated anomalous scatterer content would change proportionally.

Table 1
Number of entries for each of the anomalous scatterers present in the database.

The total number of entries is greater than 357 as some data sets are counted multiple times when they contain more than one type of anomalous scatterer.

No. of entries	Atom type
196	Se
58	Zn
26	I
22	S
12	Ca
9	Hg
7	Br
5	Fe
4	Au
3	Cd
2	Ag, Mn
1	As, Ba, Co, Cu, Pt, Rb, Ta, Tb, V, W, Yb

5.5. Analysis of the effect of radiation damage

To test the hypothesis that positive correlations between the measurement errors for members of a Bijvoet pair can arise from the effects of radiation damage, we searched the Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC; Grabowski *et al.*, 2016; http://proteindiffraction.org/) to find a data set with the keyword 'SAD', strong anomalous signal and high redundancy so that subsets of the full data could be analysed. The search yielded the data for PDB entry 3ot2 (Joint Center for Structural Genomics, unpublished work) with accession identifier https://doi.org/10.18430/M33OT2. The data set comprises 360 images, which were integrated using *XDS* (Kabsch, 2010) from *XDSGUI* (https://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/XDSGUI). Subsets of the integrated data were scaled and merged in the same package before comparing the values obtained for the error-correlation parameter as a function of resolution.

All calculations were performed on a Dell Precision 5820 machine with 128 GB RAM and an Intel Xeon W-2145 CPU @ 3.7 GHz × 16, running the CentOS version 7 operating system.

6. Results

6.1. Overview of the curated database

The curated database consisted of 357 data sets for crystals representing a total of 23 different anomalous scatterers (Table 1). In 22 cases, a mixture of anomalous scatterer types contribute strongly (with secondary anomalous scatterers contributing up to 50% of the total anomalous scattering). The space-group sampling of the database is similar to the space-group sampling of the wwPDB (Wukovitz & Yeates, 1995). Of the 357 data sets, 251 had intensity data deposited, while the rest had structure-factor data alone. Fig. 3 shows distributions for a number of characteristics of the data. The resolution of the data sets ranges from 0.93 to 3.6 Å, with the total anomalous scattering ranging from the equivalent of 0.05 to 134 fully occupied atoms. The database included data collected across a range of wavelengths from 0.81 to 2.29 Å; the largest peak in the distribution of wavelengths includes 143 Se-SAD data sets collected near the Se absorption edge at about 0.98 Å. There are three other notable peaks in the wavelength distribution: one near 0.9 Å, largely corresponding to high-energy remote Se data, one near 1.3 Å, corresponding to the Zn absorption edge, and one at 1.5418 Å, corresponding to Cu $K\alpha$ home X-ray sources. The map-to-model correlations range from values of around 0.2 to up to 0.8 for data sets with very high anomalous signal.

6.2. Estimation of the total anomalous scattering

The SCA mode estimates the scattering power of the anomalous substructure, measured in terms of the equivalent number of fully occupied primary anomalous scatterers, as discussed in Section 3. The estimated number correlates well with the total anomalous scattering determined from the reference substructure, with a log–log correlation coefficient of 0.72 for data deposited as intensities (Fig. 4). (Supplementary Fig. S1 shows an equivalent plot also including data deposited as amplitudes; the correlation coefficient is still 0.72 but there are more outliers, likely reflecting the difficulty in reversing the transformation from intensities to amplitudes.) The estimates are also consistent across different element types (see Supplementary Fig. S2). However, the total anomalous scattering tends to be slightly underestimated (or, alternatively, the refined occupancies could be slightly overestimated).

6.3. Effects of radiation damage

PDB entry 3ot2 belongs to the cubic space group *P*23, and the diffraction data deposited in the IRRMC comprise 360 images with 0.5° oscillation per image, giving a total of 180° of data. With the high symmetry, there is greater than tenfold average redundancy for each observation of the plus or minus hand of the Bijvoet pairs. To confirm the presence of radiation damage during data collection, a model-phased difference Fourier was computed, comparing the data processed from the first 90 images with those from the last 90 images. The strongest peaks in the resulting map reveal the decarboxylation of a number of acidic side chains, a cluster of which are shown in Fig. 5.

The diffraction data were reprocessed to include four progressively wider ranges of radiation dose, including the first 90, 180 or 270 or all 360 images. The substructure content analysis was carried out for each merged data set, comparing the values of $\rho_{\pm}$ obtained in each analysis. As expected from the hypothesis that a correlation of errors between Bijvoet mates can arise from merging data suffering from different degrees of radiation damage, the values of $\rho_{\pm}$ increase with both resolution and total radiation dose (Fig. 6). The overall mean values of $\rho_{\pm}$ are 0.086 for data from the first 90 images, 0.122 for the first 180 images, 0.149 for the first 270 images and 0.160 for all 360 images.

7. Discussion

In SAD phasing based on structure-factor amplitudes, the difficulty of reliably extracting the anomalous signal from the

noise introduced by intensity-measurement errors is further complicated by difficulties in converting intensity errors into amplitude errors. Our experiences with accounting for the effect of intensity-measurement errors in molecular replacement (Read & McCoy, 2016) suggested that the effects of scalar errors in intensity measurements could be approximated well as complex errors in structure factors, transforming the intensity data into effective amplitudes (F_e^+ and F_e^-) and Luzzati-type weighting parameters (D_O^+ and D_O^-). Numerical tests showed that the joint distribution of the true amplitudes in the Bijvoet pair, given the observed intensities, was approximated extremely well by this treatment when the measurement errors in the Bijvoet pair are independent. However, the results of preliminary test calculations suggested that in fact measurement errors are positively correlated. A measurement error-correlation parameter, $\rho_{\pm}$, was introduced and further numerical tests showed that the joint distribution of the true amplitudes could still be approximated extremely well, even with strongly correlated measurement errors. This error treatment, therefore, will underlie our continuing work on an intensity-based SAD likelihood target, termed iSAD, which should strengthen the use of SAD data sets with marginal signal to noise.

The joint distributions of Bijvoet mates require knowledge of the atomic composition of the crystal and the atomic scattering factors (including the anomalous, or imaginary, contributions), which is generally only known approximately when collecting diffraction data from a crystal. The role of atomic composition in anomalous scattering can be summarized by a complex correlation parameter, ρ_{FF}, which varies smoothly with resolution and can therefore be determined in resolution shells. The joint distribution of observed amplitudes that takes account of the effects of anomalous scattering (ρ_{FF}) and measurement error (F_e^+, D_O^+ , F_e^- and D_O^-), as well as the

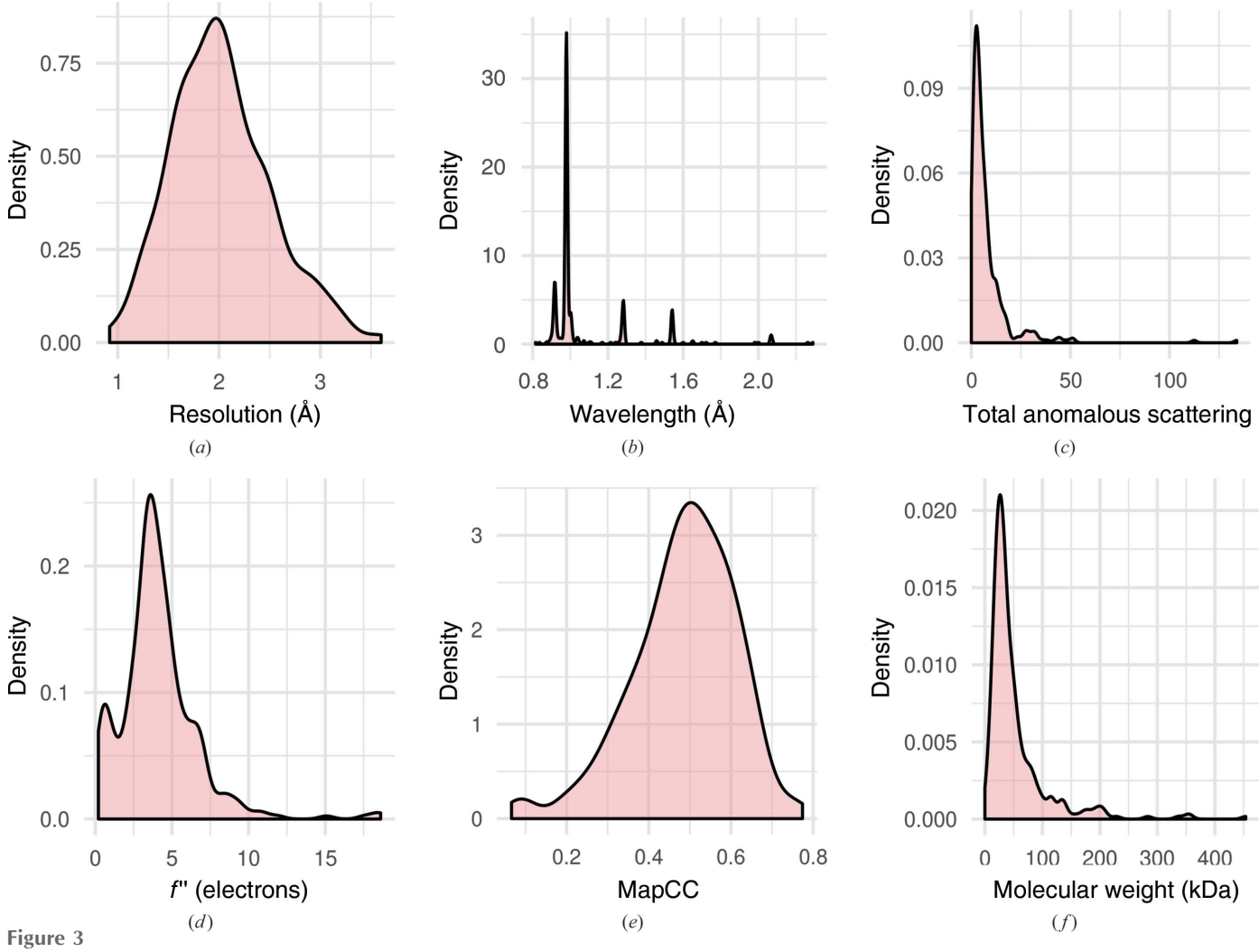


Figure 3
Distributions of relevant characteristics of data sets in the database. The vertical axes represent kernel density distribution. (*a*) Distribution of resolution limits; data to worse than 4 Å resolution were excluded. (*b*) Distribution of wavelengths. (*c*) Distribution of total anomalous scattering for the reference substructures, corresponding to the number of fully occupied anomalous scatterers with equivalent scattering power. This is measured as the f''-weighted sum of squared occupancies of refined sites to account for both primary and secondary anomalous scatterers. (*d*) Distribution of refined f'' values after the log-likelihood-gradient completion protocol. (*e*) Distribution of correlation coefficients between the experimentally phased map at the end of the log-likelihood-gradient completion protocol and density corresponding to the deposited model. (*f*) Distribution of molecular weights of the target proteins.

correlations in measurement errors between Bijvoet pairs ($\rho_{\pm}$), is the basis for a likelihood target that can be optimized in terms of the two types of correlation parameter, ρ_{FF} and $\rho_{\pm}$. Given the atomic composition of the protein component of the crystal, as well as the presumed identity of the primary anomalous scatterer, the variation of ρ_{FF} with resolution can be interpreted in terms of the content of the primary anomalous scatterer (the equivalent number of fully occupied atoms) and the average difference between the B factors of the anomalous scatterers and of other atoms in the crystal. In practice, if different hypotheses about the number of copies of the protein in the asymmetric unit were being tested, the estimated anomalous scatterer content would change proportionally.

The validity of the likelihood target and the deductions that it allows about the anomalous scatterer content were tested by carrying out calculations on our extensive curated database. This demonstrated an excellent correlation between the predicted anomalous scatterer content and the content obtained by refining the known substructures against the same data.

The results presented here demonstrate the accuracy of the new statistical model for the effects of measurement error and atomic composition on the measurement of Bijvoet pairs of reflections. The deduced anomalous scatterer content can inform strategic decisions about whether it is likely that the substructure can be determined, how difficult the problem will be (as it depends strongly on the number of atoms to be found) and how to approach the substructure determination. The success of the statistical approach depends on the quality of the measurement-error estimates; our results imply that these error estimates, at least for data used successfully for SAD phasing, are reasonably accurate.

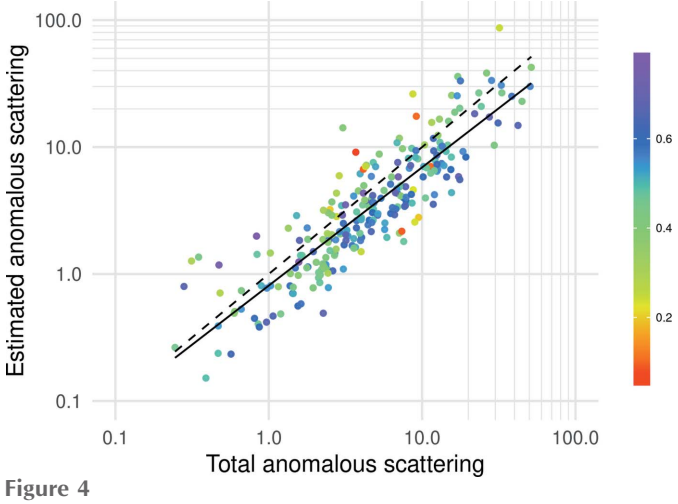


Figure 4
Estimation of the equivalent fully occupied number of primary anomalous scatterers for data deposited as intensities. The horizontal axis is the total anomalous scattering power of the gold-standard substructure (weighted sum of squared occupancies of refined sites) and the vertical axis is the estimated anomalous scattering power. The dashed black line represents a perfect prediction, while the black line shows the least-squares linear fit of the estimates. Each data point is coloured by the map correlation coefficient as shown in the legend. Both axes are plotted on a $\log_{10}$ scale.

Work in progress will build on what is presented here, showing that the results of the substructure content analysis can subsequently be used to calculate a number of measures of signal for SAD phasing: the extra information content gained by measuring Bijvoet pairs and expected values for the log-likelihood gain, figures of merit and map correlations that will be achieved in phasing once a substructure has been determined. In the longer term, we plan to implement a new iSAD phasing calculation, which should yield better quality phase information for data with low signal.

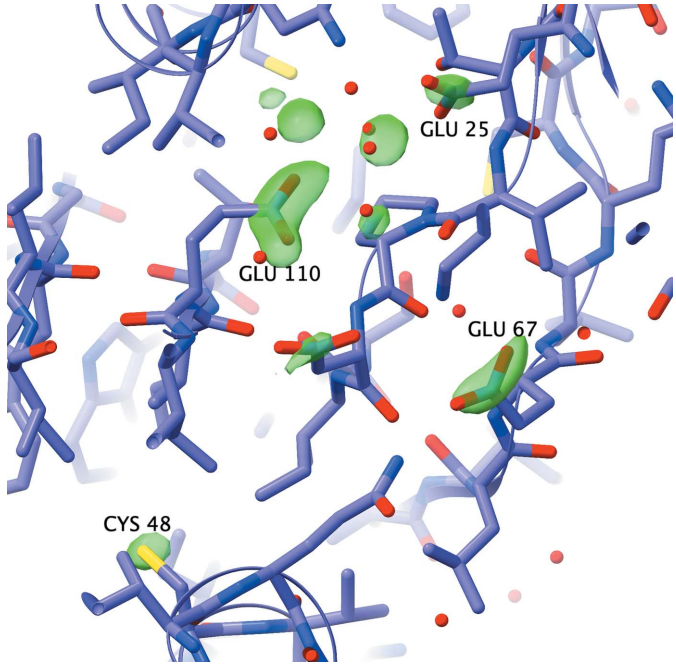


Figure 5
Difference map between data from the first quarter and the last quarter of data collection for PDB entry 3ot2, computed with map coefficients $F_{o,first} - F_{o,last}$, α_{calc}, with phases calculated from the deposited structure. The map is contoured at five times its r.m.s. value; the strongest peak, at residue Glu110 of chain A, has a height of 8.41 times the r.m.s. value. This figure was made with *ChimeraX* (Goddard *et al.*, 2018).

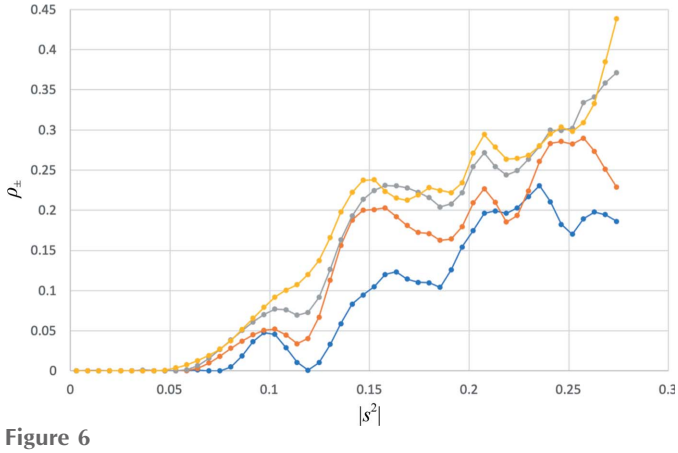


Figure 6
Error-correlation parameter, $\rho_{\pm}$, as a function of resolution for different total levels of radiation dose. The curves show values obtained from analysing data merging the first 90 images (blue), the first 180 images (orange), the first 270 images (grey) and all 360 images (yellow).

APPENDIX *A*
Conditional probability of true amplitudes given the observed intensities

The first step in deriving the desired conditional probability distribution is to obtain the joint prior distribution of true normalized amplitudes, starting from the joint distribution of normalized structure factors in (4). A change of variables to amplitudes and phases yields (15) where, for notational simplicity, Z denotes the square of the corresponding E value [*e.g.* $Z^- = (E^-)^2$] and the phase of $\mathbf{E}^{-*}$ is referred to as α^-.

$$\mathrm{p}(E^+, \alpha^+, E^-, \alpha^-) = \frac{E^+E^-}{\pi^2(1-\rho_{FF}^2)} \times \exp\left[-\frac{Z^+ + Z^- - 2\rho_{FF}E^+E^-\cos(\alpha^+ - \alpha^- - \alpha_{FF})}{1-\rho_{FF}^2}\right]. \tag{15}$$

In (15), the phase of ρ_{FF} has been ignored; if it were included, it would simply add a phase shift to the phase difference and would therefore have no effect on the integral over all phases in the next step to obtain the joint distribution of normalized amplitudes in (16).

$$\mathrm{p}(E^+, E^-) = \frac{4E^+E^-}{1-\rho_{FF}^2}\exp\left(-\frac{Z^+ + Z^-}{1-\rho_{FF}^2}\right)I_0\left(\frac{2\rho_{FF}E^+E^-}{1-\rho_{FF}^2}\right). \tag{16}$$

A change of random variables from normalized amplitudes gives the prior joint distribution of the normalized intensities in (17) (noting that, for example, $\mathrm{d}Z^- = 2E^-\mathrm{d}E^-$).

$$\mathrm{p}(Z^+, Z^-) = \frac{1}{1-\rho_{FF}^2}\exp\left(-\frac{Z^+ + Z^-}{1-\rho_{FF}^2}\right)I_0\left[\frac{2\rho_{FF}(Z^+Z^-)^{1/2}}{1-\rho_{FF}^2}\right]. \tag{17}$$

Assuming that the measured normalized intensities are related to the true values by the addition of correlated measurement errors drawn from a bivariate normal distribution, the conditional distribution of the observed normalized intensities is given in (5), which is repeated here for convenience.

$$\mathrm{p}(Z_\mathrm{O}^+, Z_\mathrm{O}^-; Z^+, Z^-) = \frac{1}{2\pi[(1-\rho_\pm^2)\sigma_{Z_\mathrm{O}^+}^2\sigma_{Z_\mathrm{O}^-}^2]^{1/2}} \times \exp\left[-\frac{(Z_\mathrm{O}^+ - Z^+)^2}{2\sigma_{Z_\mathrm{O}^+}^2(1-\rho_\pm^2)} - \frac{(Z_\mathrm{O}^- - Z^-)^2}{2\sigma_{Z_\mathrm{O}^-}^2(1-\rho_\pm^2)} + \frac{\rho_\pm(Z_\mathrm{O}^+ - Z^+)(Z_\mathrm{O}^- - Z^-)}{\sigma_{Z_\mathrm{O}^+}\sigma_{Z_\mathrm{O}^-}(1-\rho_\pm^2)}\right]. \tag{5}$$

The joint probability of both pairs of true and observed intensities, $\mathrm{p}(Z_\mathrm{O}^+, Z_\mathrm{O}^-, Z^+, Z^-)$, is obtained by multiplying together the expressions in (17) and (5), and the probability distribution of the observed pair of intensities is then obtained by integrating over all possible values of the true intensities, shown in (18).

$$\mathrm{p}(Z_\mathrm{O}^+, Z_\mathrm{O}^-) = \iint_0^\infty \mathrm{p}(Z_\mathrm{O}^+, Z_\mathrm{O}^-; Z^+, Z^-)\mathrm{p}(Z^+, Z^-)\,\mathrm{d}Z^+\,\mathrm{d}Z^-. \tag{18}$$

Finally, Bayes' theorem is used to obtain the conditional probability of the true pair of intensities given the observed pair from the expressions in (5), (17) and (18), and a change of variables gives the probability distribution for normalized amplitudes shown in (19). As above, for notational simplicity, Z is used for the square of the corresponding E value.

$$\mathrm{p}(E^+, E^-; Z_\mathrm{O}^+, Z_\mathrm{O}^-) = 4E^+E^-\frac{\mathrm{p}(Z_\mathrm{O}^+, Z_\mathrm{O}^-; Z^+, Z^-)\mathrm{p}(Z^+, Z^-)}{\mathrm{p}(Z_\mathrm{O}^+, Z_\mathrm{O}^-)}. \tag{19}$$

In the evaluation of (19) used for numerical tests, the double integral from (18) is carried out analytically in *Mathematica* (version 12.0; Wolfram Research).

APPENDIX *B*
Conditional probability of true amplitudes from the iSAD approximation

The first step in developing the desired probability distribution is to construct the joint distribution of the true normalized structure factors along with the phased structure factors corresponding to the effective amplitudes. The mathematical form for the relationships among these structure factors is the same as that considered for phasing SAD data when there are calculated structure factors from a substructure model, so the derivations below follow a similar outline to previous work on the SAD likelihood target (McCoy *et al.*, 2004). To define the distribution, we need four new complex covariances (equivalent to complex correlations, because the variables are normalized), given in (20).

$$\langle \mathbf{E}^+\mathbf{E}_\mathrm{e}^{+*}\rangle = \langle \mathbf{E}^+(D_\mathrm{O}^+\mathbf{E}^+ + \boldsymbol{\Delta}^+)^*\rangle = D_\mathrm{O}^+, \tag{20a}$$

$$\langle \mathbf{E}^-\mathbf{E}_\mathrm{e}^{-*}\rangle = \langle \mathbf{E}^-(D_\mathrm{O}^-\mathbf{E}^- + \boldsymbol{\Delta}^-)^*\rangle = D_\mathrm{O}^-, \tag{20b}$$

$$\langle \mathbf{E}^+\mathbf{E}_\mathrm{e}^{-}\rangle = \langle \mathbf{E}^+(D_\mathrm{O}^-\mathbf{E}^- + \boldsymbol{\Delta}^-)\rangle = D_\mathrm{O}^-\rho_{FF}, \tag{20c}$$

$$\langle \mathbf{E}^-\mathbf{E}_\mathrm{e}^{+}\rangle^* = \langle \mathbf{E}^-(D_\mathrm{O}^+\mathbf{E}^+ + \boldsymbol{\Delta}^+)\rangle^* = D_\mathrm{O}^+\rho_{FF}^*. \tag{20d}$$

The overall joint distribution of these four complex structure factors is a multivariate complex normal distribution (21*a*) in which the expected values (before any measurements or other information) are all zero and the covariance matrix is given in (21*b*).

$$\mathrm{p}(\mathbf{E}^+, \mathbf{E}^{-*}, \mathbf{E}_\mathrm{e}^+, \mathbf{E}_\mathrm{e}^{-*}) = \frac{1}{|\pi\Sigma|}\exp\left[-\begin{pmatrix}\mathbf{E}^+\\ \mathbf{E}^{-*}\\ \mathbf{E}_\mathrm{e}^+\\ \mathbf{E}_\mathrm{e}^{-*}\end{pmatrix}^H \Sigma^{-1}\begin{pmatrix}\mathbf{E}^+\\ \mathbf{E}^{-*}\\ \mathbf{E}_\mathrm{e}^+\\ \mathbf{E}_\mathrm{e}^{-*}\end{pmatrix}\right], \tag{21a}$$

$$\text{where } \Sigma = \begin{pmatrix} 1 & \rho_{FF} & D_\mathrm{O}^+ & D_\mathrm{O}^-\rho_{FF} \\ \rho_{FF}^* & 1 & D_\mathrm{O}^+\rho_{FF}^* & D_\mathrm{O}^- \\ D_\mathrm{O}^+ & D_\mathrm{O}^+\rho_{FF} & 1 & \rho_{FF,\mathrm{obs}} \\ D_\mathrm{O}^-\rho_{FF}^* & D_\mathrm{O}^- & \rho_{FF,\mathrm{obs}}^* & 1 \end{pmatrix}. \tag{21b}$$

The basic strategy to obtain the desired conditional distribution is to change variables to amplitudes and phases, integrate over all the unknown phases to obtain a joint distribution of the amplitudes, and then apply Bayes' theorem. One route to this result is given in (22).

$$\mathrm{p}(E^+, E^-; E_\mathrm{e}^+, E_\mathrm{e}^-) = \iiint_0^{2\pi} \frac{\mathrm{p}(E^+;E^-,\alpha^-,E_\mathrm{e}^+,\alpha_\mathrm{e}^+,E_\mathrm{e}^-,\alpha_\mathrm{e}^-)\mathrm{p}(E^-,\alpha^-,E_\mathrm{e}^+,\alpha_\mathrm{e}^+,E_\mathrm{e}^-,\alpha_\mathrm{e}^-)}{\mathrm{p}(E_\mathrm{e}^+,E_\mathrm{e}^-)}\,\mathrm{d}\alpha^-\,\mathrm{d}\alpha_\mathrm{e}^+\,\mathrm{d}\alpha_\mathrm{e}^-. \quad (22)$$

In the triple integral, all of the phase terms contain phase differences so, if one phase is fixed at an arbitrary value, the others will vary over all possible values relative to each other and to the fixed phase. For this reason, one of the phase integrals can be omitted and the remaining double integral can simply be multiplied by 2π, as shown in (23).

$$\mathrm{p}(E^+, E^-; E_\mathrm{e}^+, E_\mathrm{e}^-) = 2\pi \iint_0^{2\pi} \frac{\mathrm{p}(E^+;E^-,0,E_\mathrm{e}^+,\alpha_\mathrm{e}^+,E_\mathrm{e}^-,\alpha_\mathrm{e}^-)\mathrm{p}(E^-,0,E_\mathrm{e}^+,\alpha_\mathrm{e}^+,E_\mathrm{e}^-,\alpha_\mathrm{e}^-)}{\mathrm{p}(E_\mathrm{e}^+,E_\mathrm{e}^-)}\,\mathrm{d}\alpha_\mathrm{e}^+\,\mathrm{d}\alpha_\mathrm{e}^-. \quad (23)$$

For a similar reason, the phases of ρ_{FF} and $\rho_{FF,\mathrm{obs}}$ are ignored because they would just add a constant offset to the phase differences. The double integral is carried out analytically in *Mathematica* for the numerical tests. The three probability distributions needed for (23) are provided below.

The joint probability distribution for three phased structure factors, $\mathrm{p}(E^-, \alpha^-, E_\mathrm{e}^+, \alpha_\mathrm{e}^+, E_\mathrm{e}^-, \alpha_\mathrm{e}^-)$, is obtained by analogy to (21), but omitting $\mathbf{E}^+$ as well as the first row and column of the covariance matrix and then changing the complex variables to amplitude and phase variables.

The probability of the amplitude of E^+ given the other three phased structure factors is obtained by first partitioning the covariance matrix from (21) to obtain the conditional distribution of $\mathbf{E}^+$ in (24).

$$\mathrm{p}(\mathbf{E}^+; \mathbf{E}^{-*}, \mathbf{E}_\mathrm{e}^+, \mathbf{E}_\mathrm{e}^{-*}) = \frac{1}{\pi\Sigma}\exp[(\mathbf{E}^+ - \langle\mathbf{E}^+\rangle)^*\Sigma^{-1}(\mathbf{E}^+ - \langle\mathbf{E}^+\rangle)], \quad (24)$$

$$\text{where } \langle\mathbf{E}^+\rangle = \Sigma_{12}\Sigma_{22}^{-1}\begin{pmatrix}\mathbf{E}^{-*},\\ \mathbf{E}_\mathrm{e}^+\\ \mathbf{E}_\mathrm{e}^{-*}\end{pmatrix},$$

$$\Sigma = 1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\Sigma_{12} = \begin{pmatrix}\rho_{FF} & D_\mathrm{O}^+ & D_\mathrm{O}^-\rho_{FF}\end{pmatrix},$$

$$\Sigma_{21} = \Sigma_{12}^T,$$

$$\Sigma_{22} = \begin{pmatrix}1 & D_\mathrm{O}^+\rho_{FF} & D_\mathrm{O}^-,\\ D_\mathrm{O}^+\rho_{FF} & 1 & \rho_{FF,\mathrm{obs}},\\ D_\mathrm{O}^- & \rho_{FF,\mathrm{obs}} & 1\end{pmatrix}.$$

Next, after a change of variables from the complex $\mathbf{E}^+$ to its amplitude (E^+) and phase (α^+), the expression in (24) is integrated over all possible values of α^+. This integral has an analytical solution, given in (25).

$$\mathrm{p}(E^+; \mathbf{E}^{-*}, \mathbf{E}_\mathrm{e}^+, \mathbf{E}_\mathrm{e}^{-*}) = \frac{2E^+}{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}\exp\left(-\frac{E^{+2} + |\langle\mathbf{E}^+\rangle|^2}{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}\right) \times I_0\left(\frac{2E^+|\langle\mathbf{E}^+\rangle|}{1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}\right). \quad (25)$$

Finally, the prior joint probability distribution of the effective amplitudes, $\mathrm{p}(E_\mathrm{e}^+, E_\mathrm{e}^-)$, as given in (9) above, is repeated here for convenience.

$$\mathrm{p}(E_\mathrm{e}^+, E_\mathrm{e}^-) = \frac{4E_\mathrm{e}^+E_\mathrm{e}^-}{1 - |\rho_{FF,\mathrm{obs}}|^2}\exp\left(-\frac{E_\mathrm{e}^{+2} + E_\mathrm{e}^{-2}}{1 - |\rho_{FF,\mathrm{obs}}|^2}\right) \times I_0\left(\frac{2|\rho_{FF,\mathrm{obs}}|E_\mathrm{e}^+E_\mathrm{e}^-}{1 - |\rho_{FF,\mathrm{obs}}|^2}.\right) \quad (9)$$

Acknowledgements

We thank Tom Terwilliger and Zbyszek Dauter for kindly sharing diffraction data sets used in this study.

Funding information

This research was supported by funding from CCP4 (KSH), a Wellcome Trust Principal Research Fellowship (RJR: grant 209407/Z/17/Z) and the NIH (grant P01GM063210 to RJR), which is gratefully acknowledged.

References

Banumathi, S., Zwart, P. H., Ramagopal, U. A., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1085–1093.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bunkóczi, G., McCoy, A. J., Echols, N., Grosse-Kunstleve, R. W., Adams, P. D., Holton, J. M., Read, R. J. & Terwilliger, T. C. (2015). *Nat. Methods*, **12**, 127–130.
Dauter, Z., Dauter, M. & Dodson, E. J. (2002). *Acta Cryst.* D**58**, 494–506.
Evans, G. & Pettifer, R. F. (2001). *J. Appl. Cryst.* **34**, 82–86.
Garcia-Bonete, M.-J. & Katona, G. (2019). *Acta Cryst.* A**75**, 851–860.
Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. (2018). *Protein Sci.* **27**, 14–25.
Grabowski, M., Langner, K. M., Cymborowski, M., Porebski, P. J., Sroka, P., Zheng, H., Cooper, D. R., Zimmerman, M. D., Elsliger, M.-A., Burley, S. K. & Minor, W. (2016). *Acta Cryst.* D**72**, 1181–1193.
Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* D**59**, 1974–1977.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Cryst.* D**75**, 861–877.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
McCoy, A. J. & Read, R. J. (2010). *Acta Cryst.* D**66**, 458–469.
McCoy, A. J., Stockwell, D. H., Sammito, M. D., Oeffner, R. D., Hatti, K. S., Croll, T. I. & Read, R. J. (2021). *Acta Cryst.* D**77**, 1–10.
McCoy, A. J., Storoni, L. C. & Read, R. J. (2004). *Acta Cryst.* D**60**, 1220–1228.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Read, R. J., Adams, P. D. & McCoy, A. J. (2013). *Acta Cryst.* D**69**, 176–183.
Read, R. J. & McCoy, A. J. (2011). *Acta Cryst.* D**67**, 338–344.
Read, R. J. & McCoy, A. J. (2016). *Acta Cryst.* D**72**, 375–387.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Acta Cryst.* D**65**, 582–601.
Terwilliger, T. C., Bunkóczi, G., Hung, L.-W., Zwart, P. H., Smith, J. L., Akey, D. L. & Adams, P. D. (2016*a*). *Acta Cryst.* D**72**, 346–358.
Terwilliger, T. C., Bunkóczi, G., Hung, L.-W., Zwart, P. H., Smith, J. L., Akey, D. L. & Adams, P. D. (2016*b*). *Acta Cryst.* D**72**, 359–374.
Wang, J., Dauter, M. & Dauter, Z. (2006). *Acta Cryst.* D**62**, 1475–1483.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Wukovitz, S. W. & Yeates, T. O. (1995). *Nat. Struct. Mol. Biol.* **2**, 1062–1067.